Estimating the size of the solution space of metabolic networks

Alfredo Braunstein[1,2], Roberto Mulet[2], Andrea Pagnani[3]

[1]ISI Foundation Viale Settimio Severo 65, Villa Gualino, I-10133 Torino, Italy
[2]"Henri-Poincaré-Group" of Complex Systems and Department of Theoretical Physics, Physics Faculty, University of Havana, La Habana, CP 10400, Cuba
[3]Politecnico di Torino, Corso Duca degli Abruzzi 34, I-10129, Torino, Italy

Email: Alfredo Braunstein - alfredo.braunstein@polito.it; Roberto Mulet*- mulet@fisica.uh.cu; Andrea Pagnani*- pagnani@isi.it;

*Corresponding author

Abstract

Background: Cellular metabolism is one of the most investigated system of biological interactions. While the topological nature of individual reactions and pathways in the network is quite well understood there is still a lack of comprehension regarding the global functional behavior of the system. To get some insight in this direction, in the last few years powerful theoretical methods such as extreme pathways calculation and flux-balance analysis (FBA) have been introduced. These methods strongly rely on the hypothesis that the organism maximizes an objective function. However only under very specific biological conditions (*e.g.* maximization of biomass for E-Coli in reach nutrient medium) the cell seems to obey such optimization law. A more refined analysis not assuming extremization remains an elusive task for large metabolic systems due to algorithmic constraints.

Results: In this work we propose a novel algorithmic strategy that allows for an efficient characterization of the whole set of stable fluxes compatible with the metabolic constraints. Using a technique derived from the fields of statistical physics and information theory we designed a stochastic algorithm to estimate the size of the affine space containing all possible steady-state flux distributions of metabolic networks. The algorithm, based on the well-known Bethe approximation, allows the computation in polynomial time of the volume of a non full-dimensional convex polytope in high dimensions. We show results for toy models that compare very well with the results of exact algorithms. The result of our algorithm match closely the prediction of Monte Carlo based estimations of the flux distributions of the Red Blood Cell metabolic network. It is then used to analyze the statistical properties of the average fluxes of the reactions in the E-Coli metabolic network and finally to test the effect of gene knock-outs on the size of the solution space of the E-Coli central metabolism.

Conclusions: We propose a novel efficient algorithmic strategy to estimate the size and shape of the affine space of a non full-dimensional convex polytope in high dimensions. The method is shown to obtain, quantitatively and qualitatively compatible results compared with known algorithms (where this comparison is possible) being still efficient on the analysis of large biological systems, where all other known strategies experience an explosion in algorithmic time.

Background

Cellular metabolism is a complex biological problem. It can be viewed as a chemical engine that transforms available raw materials into energy or into the building blocks needed for the biological function of the cells. In more specific terms a metabolic network is indeed a processing system transforming *input* metabolites (nutrients), into output metabolites (amino acids, lipids, sugars etc.) according to very strict molecular proportions, often referred as stoichiometric coefficients of the reactions.

Although the general topological properties of these networks are well characterized, see for example [1–3], and non-trivial pathways are well known for many species [4] the cooperative role of these pathways is hard to comprehend. In fact, the large sizes of these networks, usually containing hundreds of metabolites and even more reactions, makes the comprehension of the principles that govern their global function a challenging task. Therefore, a necessary step to achieve this goal is the use of mathematical models and the development of novel statistical techniques to characterize and simulate these networks.

For example, it is well known that under evolutionary pressure, prokaryotes cells like E-Coli behave optimizing their growth performance [5]. Flux Balance Analysis (FBA) provides a powerful tool to predict, from the whole space of phenotipic states, which one will these cells acquire. In few words one may say that FBA maximizes a linear function (usually the growth rate of the cell) subject to biochemical and thermodynamic constraints [6]. On the other hand, cells with genetically engineered knockouts or bacterial strains that were not exposed to evolution pressures, need not to optimize their growth. In fact, the method of minimization metabolic adjustment (MOMA) [7] has shown that knockout metabolic fluxes undergo a minimal redistribution with respect to the flux configuration of the wild type. Yet, in more general situations, the results are unpredictable, therefore, a tool to characterize the shape and volume of the whole space of possible phenotipic solutions must be welcome.

Unfortunately, this characterization has remained an elusive task. As far as we know the attempts to obtain such a characterization were always based on the Monte Carlo sampling of the steady-state flux space [8]. But unfortunately, it appeared that this kind of sampling is a very expensive calculation and is unsuitable for very large networks. To address this problem we propose, using a technique derived from the fields of statistical physics and information theory, an algorithm that may efficiently characterize the whole set of stable fluxes compatibles with the stoichiometric constraints.

Mathematical Model

As already mentioned, a metabolic network is an engine that converts metabolites into other metabolites through a series of intra-cellular intermediate steps. The fundamental equation characterizing all functional states of a reconstructed biochemical reaction network is a mass conservation law that imposes simple linear constraints between the incoming and outcoming fluxes at any chemical reaction. It is call the *dynamic mass balance equation:*

$$\frac{\partial \boldsymbol{\rho}}{\partial t} = \mathbf{i} + \hat{S} \cdot \boldsymbol{\nu} - \mathbf{o} \tag{1}$$

where $\boldsymbol{\rho}$ is the vector of the M metabolite concentrations in the network. $\mathbf{i}$ ($\mathbf{o}$) is the input (output) vector of fluxes, and $\boldsymbol{\nu}$ are the reaction fluxes governed by the $M \times N$ stoichiometric linear operator $\hat{S}$ (usually named stoichiometric matrix) encoding the coefficient of the M intra-cellular relations among the N fluxes.

As long as just steady-state cellular properties are concerned one can assume that a variation in the concentration of metabolites in a cell can be ignored and considered as constant. Therefore in case of fixed external conditions one can assume metabolites (quasi) stationarity and consequently the *lhs* of 1 can be set to zero. Under these hypotheses the problem of finding the metabolic fluxes compatible with flux-balance is mathematically described by the linear system of equations

$$\hat{S} \cdot \boldsymbol{\nu} = \mathbf{o} - \mathbf{i} \equiv \mathbf{b} \tag{2}$$

where $\mathbf{b}$ is the net metabolite uptake by the cell. Without loss of generality we can assume that the stoichiometric matrix $\hat{S}$ has full rows rank, *i.e.* that $\mathrm{rank}(\hat{S}) = M$, since linearly dependent equations can be easily identified and removed. Knowing that the number of metabolites M is lower than the number of fluxes N the subspace of solutions is a $(N-M)$-dimensional manifold embedded in the N-dimensional space of fluxes. In addition, the positivity of fluxes, together with the experimentally

accessible values for the maximal fluxes, limit further the space of feasible solutions. This fact may be expressed by the following inequalities:

$$\mathbf{0} \leq \boldsymbol{\nu} \leq \boldsymbol{\nu}^{\mathbf{max}} \tag{3}$$

in such a way that together, 2 and 3, define the convex set of all the allowed time-independent phenotipic states of a given metabolic network.

Sub-dimensional volumes

Mathematically speaking, the space of feasible solutions consistent with the equations 2 constitutes an affine space $V \subset \mathbb{R}^N$ of dimension $N-M$. The set of inequalities 3 then defines a convex polytope $\Pi \subset V$ that, from the metabolic point of view, may be considered as the allowed configuration space for the cell states. With the scope of describing it, we will be interested in computing the volume of this space and certain volumes of subspaces of it. Although conceptually simple, the notion of sub-dimensional volume like that of Π requires some new definitions.

Consider any linear parameterization $\phi : \mathbb{R}^{N-M} \to V \subset \mathbb{R}^N$. A popular choice for ϕ is, for instance, the inverse of the so called *lexicographical* projection *i.e.*, the projection over the first $N-M$ coordinates such that its restriction to V has an inverse. Being ϕ linear, the $(N-M) \times N$ Jacobian matrix $\hat{\lambda}$ is constant and coincides with the matrix of ϕ in the canonical bases. Denoting $\lambda = \det(\hat{\lambda}^\dagger \hat{\lambda})^{\frac{1}{2}}$, the Euclidean metric in $\mathbb{R}^N$ induces a measure on V (which does not depend on ϕ):

$$\int_V f(\boldsymbol{\nu}) d\boldsymbol{\nu} \equiv \lambda \int f(\phi(\boldsymbol{u})) d\boldsymbol{u} \tag{4}$$

allowing to compute the volume of our polytope

$$\mathbf{vol}_V(\Pi) \equiv \int_V \mathbf{1}_\Pi(\boldsymbol{\nu}) d\boldsymbol{\nu} = \lambda \int \mathbf{1}_{\phi^{-1}(\Pi)}(\boldsymbol{u}) d\boldsymbol{u} \tag{5}$$

where $\mathbf{1}_\Pi(\cdot)$ is the indicator function of the set Π. It is worth pointing out that given the linear structure of the metabolic equations, the determinant of the mapping is a (scalar) constant.

Probabilistic framework

The problem of describing the polytope Π can formulated in a probabilistic framework. We define the probability density $\mathcal{P}$ as:

$$\mathcal{P}(\boldsymbol{\nu}) = \mathbf{vol}_V(\Pi)^{-1} \mathbf{1}_\Pi(\boldsymbol{\nu}) \tag{6}$$

Marginal flux probabilities over a given set of fluxes are obtained by integrating out all remaining degrees of freedom. In particular we can define single flux marginal probability densities as integrals on the affine subspace $W = V \cap \{\nu_i = \overline{\nu}\}$:

$$P_i(\nu_i) = \int_W \mathcal{P}(\boldsymbol{\nu}) \prod_{j \neq i} d\nu_j = \mathbf{vol}_W(\Pi \cap W) \tag{7}$$

i.e. the (sub dimensional) volume of the intersection between the polytope Π and the hyperplane $\{\nu_i = \overline{\nu}\}$.

Results and Discussion

Performance on low dimensional systems

In this section we will analyze the performance of our algorithm against an exact algorithm on low dimensional polytopes. Among the different packages available in the Internet, we have chosen LRS [9], a program based on the *reverse search* algorithm presented by Avis and Fukuda in [10] that can compute the volume of non-full dimensional polytopes. Actually, it computes the volume of the lexicographically smallest representation of the polytope, that for the benchmark used below, coincides with the conventional volume estimated by our algorithm.

We have devised a specific benchmark generating random diluted stoichiometric matrices at a given ratio $\alpha = M/N$ and fixed number of terms different from zero K in each of the reactions. All fluxes were constrained inside the hypercube $0 \leq \nu_i \leq 1$. As a general strategy we have calculated several random instances of the problem and measured the volume (entropy) of the polytope using the LRS and BP algorithm. In particular, we have first generated 1000 realizations of random stoichiometric matrices with $N = 12, M = 4$. Note that $N = 12$ is around the maximum that allows simulations with LRS in reasonable time (around one hour per instance). For each polytope then we have computed the two entropies S_{LRS} and S_{BP} with both algorithms, fixing the same maximum value for the discretization $q^{\max} = 1024$ for all fluxes.

In Figure 1.2 we show how the quality of the BP measure is affected by the discretization, by displaying the histogram of the relative differences $\delta S = \frac{S_{BP} - S_{LRS}}{S_{LRS}}$ with an increasing number of bins per variable $q^{\max} = 16, 64, 256, 1024$. One can see how a finer binning of messages improves the quality of the approximation, seeminlgy converging to a

single distribution of errors. It is expected that for larger N the histograms would shrink: upon increasing the number of fluxes, loops become larger and the overall topology of the graph becomes more locally tree-like, validating the hypothesis behind the Bethe approximation. Unfortunately, the huge increase of computer time experimented in the calculation of the volumes using LRS made impossible to test systems large enough to make any reasonable scaling analysis.

Finally we address the issue of the computational complexity of the algorithm which is a crucial one if one is interested in approaching real world metabolic networks whose size typically is at least 50 times the size of the largest network that can be analyzed with exact algorithms. In Figure 1.1 we display the running time of both LRS and BP as a function of the number of fluxes N. Interesting, LRS outperforms BP up to sizes $N \sim 12$ where the running time of LRS explodes exponentially while BP maintains a modest almost-linear behavior.

Distribution of fluxes in Red Blood Cell

The algorithm was used to obtain the distribution of flux values for each of the reactions of the Red Blood Cell metabolism. The maximum allowed values for the fluxes, as well as the corresponding stoichiometric matrix were extracted directly from [8]. The network contained 46 reactions and 34 metabolites. Our distributions appear in Figure 1.3 and as can be checked by direct comparison they are almost identical to those obtained with the Monte Carlo method in with the Figure 5 in [8]. However, while the Monte Carlo method appears to be quite expensive in computational resources (the authors of [8] reported one week of computer computation in a Dell Dimension 8200 to obtain their distributions) our algorithm converged to the same results in a couple of minutes of computation on a similar machine.

Analysis of gene knock-out in E-Coli

Then, we analyze the distribution function of the average fluxes of the metabolic network of the E-Coli. The network used contains, in its original format, 1035 reactions and 626 metabolites, [11]. From this network we eliminated all those reactions with fractional stoichiometric indexes, since our algorithm in its present form is unable to deal with them, and all those metabolites with connectivity larger than 40. These metabolites are cofactors of the metabolism and their elimination helps to sparse the matrix improving the convergence of the algorithm. Finally we checked for inconsistencies (i.e. metabolites that after the previous transformations are only produced or consumed), and we obtain a matrix with 1005 reactions and 560 metabolites. We ran our algorithm on this network and computed the average fluxes in each reaction. The probability distribution function (pdf) of these average values appears in Figure 1.4. As can be clearly seen the distribution is large and may be fitted with a power law with exponent close to 1, a result that compares very well with previous simulations [12, 13] in real networks and with the approach followed by Bianconi and Zecchina [14]. Moreover, it most be mentioned that a more careful analysis of the data may suggest that the distribution of averages fluxes has a richer structure. In Figure 1.5 is presented the integrated distribution function of the average fluxes of the reactions and a clear jump appears for $\nu \approx 0.5$, and smaller ones for $\nu \approx 0.4$ and $\mu \approx 0.6$. Whether these jumps are just due to normal statistical fluctuations (and are correctly smeared out in the usual binning process done to plot the pdf in log-log scale) or reflect relevant biological or structural information of the network is not yet known.

Finally, we concentrated our efforts in the study the E-Coli *central* metabolism [11]. To simplify the analysis of the results, all cofactors were removed with the exception of ATP and ADP that have been thus considered as output and input fluxes respectively. Under these conditions, the network has 40 metabolites, 74 internal reactions and 10 external fluxes that were considered irreversible following their nominal directions. The maximum fluxes of all the reactions were considered equal to 1.

We then studied the influence of knockouts on the volume of the solution space. In each simulation the values of the maximum flux was kept constant and equal to 1 for all the reactions but the one with a knockout. For each knockout we assumed a reduction of the maximum permitted flux $q^{\max}$ equal to the 0.25 of the original value. We assumed that each reaction is knockout independently, this, despite the fact that it is known that some reactions are associated with the same enzyme.

The distribution of entropy change ($\Delta S = S_{KO} - S_0$) of the solution space for the different reaction knock-outs turns out to show two possibilities: while

most of the reactions have $\Delta S > -3$ and show little and similar impact in the size of the solution space a relevant fraction of them, has $\Delta S < -3$ but with a very uniform distribution. To understand if there is any connection between these reactions with large ΔS and the structure of the network, we show in Figure 1.6 the changes in S for the different reaction knock-outs. In the $x-$axis we plot the reactions numerated and the lines indicate the values of ΔS for each reaction. We annotate also these reactions with $\Delta S < -3$. Most of them are associated with the glycolysis pathway, being more specific, with those reactions of the glycolysis process that show little redundancy in the topology of the network. Other reactions, like *FUM, ACONT, SUC* and *SUCCD1i* appear in the Krebb cycle and again show little pathway redundancy in the network.

Finally in Figure 1.7 we display the correlations between the changes in the entropy for different reaction knock-outs and the most probable values ν^* and the average values $\bar{\nu}$ of the fluxes in the wild network. As can be seen, two kinds of regimes are divided by a clear threshold at $\nu \sim 0.6$: A first one, for small fluxes, that show consistent and slowly increasing correlations between ν and ΔS, and a second one for large fluxes, where the correlations increase rapidly but with large fluctuations. The presence of this threshold can be understood noting that reactions belonging to the linear (glycolysis) an circular (Krebb cycle) pathways are, in the wild cell, *fast*-flux reactions, with average values for the fluxes larger than 0.5. Therefore once they are knock-out, the metabolic capacities of the network are highly affected.

Conclusions

We proposed a novel algorithm to estimate the size and shape of the affine space of a non full-dimensional convex polytope in high dimensions. The algorithm was tested in specific benchmark, i.e. random diluted stoichiometric matrices at a given ratio $\alpha = M/N$ and fixed number of terms different from zero K, in each of the reactions, with results that compare very well with those of exact algorithms. Moreover, we show that while the running time of exact algorithms increases more than exponentially for already moderate sizes, our algorithm keeps a polynomial behavior for sizes as large as $N = 120$. The program was run on the Red Blood Cell metabolism, showing with less computational effort, results that compare very well with those previously obtained using Monte Carlo methods. Then, we calculate the distribution of the average values of the fluxes in the metabolism of the E-Coli and present results that are consistent with those of the literature. Finally, our program was used to study the E-Coli central metabolism, and we show that, as expected, reactions with little redundancy are the ones with more impact in the size of the space of the metabolic solutions. Specifically, most of the reactions associated with the transformation of glucose in pyruvate, belong to this set, as well as some reactions in the citric cycle. In addition we show strong correlations between the characteristics of the flux distributions of the wild type network and the changes in size of the space of solutions after reaction knock-outs.

Let us conclude by noting that in principle the presented approach can be extended to deal with constraints whose functional form is more general than linear, provided that the number of variables involved in each of the constraints remains small, as in the case of inequalities enforcing the second low of thermodynamics for the considered reactions [15]. Work is in progress in this direction.

Methods

Volume computation

From a computational point of view, the problem of the exact computation of the volume of a polytope with current methods requires the enumeration of all its vertexes. The vertex enumeration problem is $\#P$-hard [16,17], but even the problem of computing the volume, given the set of all vertexes is a big computational challenge. Various algorithms exist for calculating the exact volume of a polytope from its vertexes (for a review see [18]), and many software packages are available in the Internet. Computational limitations restrict however exact algorithmic strategies to cope with polytopes in relatively few dimensions (*e.g* $N-M$ around 10 or so). To overcome such severe limitations we will introduce a very efficient approximate computational strategy that will allow us to compute the volume and the shape of the space of solutions for real-world metabolic networks.

Although the coefficient λ could be explicitely calculated, it turns out that as far as only relative volume quantities are concerned, as in the case of the

in silico flux knock-outs introduced below, this term factors out and therefore we will drop it from the rest of the computation. If one is not interested in constant prefactors, one can proceed as follows for the integration of Eq. 6. Consider the regular orthogonal grid Λ_ϵ of side ϵ partitioning $\mathbb{R}^N$. This grid maps via ϕ^{-1} into a partition Γ_ϵ of $\phi^{-1}(\Pi)$. The number of cells $\mathcal{N}_\epsilon$ of Λ_ϵ intersecting Π is equal to the numbers of cells of Γ_ϵ intersecting $\phi^{-1}(\Pi)$. Finally, the volume in Eq. 5 is proportional to $\lim_{\epsilon\to 0}\epsilon^{N-M}\mathcal{N}_\epsilon$. The same applies to the computation of the marginal of Eq. 7, now noting that the constant pre-factor does not depend on ν_i. The reader may wonder why one does not directly integrate Eq. 5 by using its right hand side, the problem is that the change of coordinates induced by ϕ may destruct the sparsity of the original matrix $s_{a,i}$, a condition that turns out to be crucial for the kind of approximation that we will use.

When dealing with integer coefficients $s_{i,a}$, as the ones appearing in stoichiometric relations, a further simplification in the approximate volume computation is possible: one can (always ingoring a constant pre-factor) restrict further to integer solutions of the system of equations. Geometrically, the space is tiled with small hypercubes and we are actually counting the number of hypercubes exactly fulfilling the stoichiometric equations. In summary, for any ϵ the computation of an ϵ-approximation of the volume has been recast into a discrete combinatorial optimization problem that can described (with a slight abuse of notation) by the same Eqs. 2 with now discrete variables $\nu_i \in \{0, 1, ..., q_i^{\max}\}$, for $q_i^{\max}$ equal to the integer part of $q^{\max} \times \nu_i^{\max}$, where the integer $q^{\max}$ is the granularity of the approximation.

Belief Propagation

The metabolic problem can be cast into a constraint satisfaction framework where the M stoichiometric relations impose a constraint onto a subset of the metabolic fluxes. Let A be the set of equations and I the set of fluxes. Consider the *a-th* row of $\hat{S}$, and let $\{i_1, \ldots, i_{n_a}\} \equiv \{i \in a\} \subset I$ be the labels of the fluxes involved in the considered equation having stoichiometric coefficients different from zero. Let also $\{a_1, \cdots, a_{n_i}\} \equiv \{a \in i\} \subset A$ be the labels of the equations in which flux i participates. The emerging structure is a bipartite graph, with two types of nodes: *variable* nodes representing the fluxes of the reactions and *factor* nodes imposing mass conservation. In this case marginals become q-modal probability densities that for large values of $q_i^{\max}$ will approximate better and better the continuous set of probabilities.

Under the hypothesis that the factor graph is a tree it can be shown [19, 20] that a given flux vector $\boldsymbol{\nu}$ satisfying all flux-balance constraints can be expressed as a product of flux and reaction marginals [20–22]:

$$\mathcal{P}(\boldsymbol{\nu}) = \prod_{a\in A} P_a(\{\nu_l\}_{l\in a}) \prod_{i\in I} P_i(\nu_i)^{1-d_i} \qquad (8)$$

where d_i is the number of equations in which flux ν_i participates (*i.e.* the degree of site i). The marginal probabilities are defined as:

$$\begin{aligned} P_i(\nu_i) &= \sum_{\{\nu_j\}_{j\neq i}} \mathcal{P}(\boldsymbol{\nu}) \\ P_a(\{\nu_l\}_{l\in a}) &= \sum_{\{\nu_j\}_{j\notin a}} \mathcal{P}(\boldsymbol{\nu}) \ . \end{aligned} \qquad (9)$$

Belief Propagation (BP) is a local iterative algorithm that allows for the computations of marginal probability distributions, which is exact on trees, and perform reasonably well on locally tree-like structures [20–23]. This approximation scheme allows the computation of the (logarithm of the) number of solutions via the entropy that can be expressed in terms of flux marginals:

$$\begin{aligned} S &\equiv -\sum_{\boldsymbol{\nu}} \mathcal{P}(\boldsymbol{\nu}) \ln \mathcal{P}(\boldsymbol{\nu}) \\ &= \sum_{a\in A} \sum_{\{\nu_j\}_{j\in a}} P_a(\{\nu_{j\in a}\}) \log P_a(\{\nu_{j\in a}\}) \\ &- \sum_{i\in I} \sum_{\nu_i} (d_i - 1) P_i(\nu_i) \log P_i(\nu_i) \end{aligned} \qquad (10)$$

One may wonder how such an approach could be useful in a *real-world* situation where the graph is not a tree. Interestingly enough, metabolic networks are sparse, *i.e.* the number of metabolites that typically participate to a certain reaction is small with respect to the number of metabolites M, moreover one can reasonably assume the typical loop length to be large enough to ensure weak statistical dependence of neighboring sites which lay at the heart of the Bethe approximation [24, 25]. The algorithm is based on two type of messages exchanged from variable nodes to functional nodes, and vice versa:

- $\mu_{i\to a}(\nu)$: the probability that flux i takes value ν in the absence of reaction a.

- $m_{a\to i}(\nu)$: the non-normalized probability that the balance in reaction a is fulfilled given that flux i takes value ν.

The two quantities satisfy the following set of functional equations:

$$\begin{aligned} m_{a\to i}(\nu_i) &= \sum_{\{\nu_l\}_{l\in a\setminus i}} \mu_{l\to a}(\nu_l)\ \delta\Big(\sum_{l\in a} s_{a,l}\nu_l\,;\,b_a\Big) \\ \mu_{i\to a}(\nu_i) &= C_{i\to a}\prod_{b\in i\setminus a} m_{b\to i}(\nu_i) \end{aligned} \tag{11}$$

where $C_{i\to a}$ is a constant enforcing the normalization of the probability $\mu_{i\to a}(\nu)$ and $\delta(\cdot\,;\cdot)$ is the Kronecker delta function. The set of equations 11 can be solved iteratively and upon convergence of the algorithm one can compute the marginal flux distributions as:

$$\begin{aligned} P_a(\{\nu_l\}_{l\in a}) &= \sum_{\{\nu_l\}_{l\in a}} \mu_{l\to a}(\nu_l)\ \delta\Big(\sum_{l\in a} s_{a,l}\nu_l\,;\,b_a\Big) \\ P_i(\nu) &= \prod_{l\in i} m_{l\to i}(\nu)\ . \end{aligned} \tag{12}$$

A brute force integration of the discrete set of equation would be much too inefficient for analyzing large networks, due to the multiple dimensional sum over $\{0,\dots,q_l^{\max}\}_{l\in a\setminus i}$ in the previous equation. A relevant speed-up in the convergence of these equations can be achieved by noting that the convolution product in 11 can be efficiently solved with a recursion. Let us relabel the set of fluxes in equation a as $\{l\in a\}\equiv\{1,\dots,n_a\}$ and let assume for the sake of clarity that $b_a=0$:

$$R^{(k+1)}(\nu)=\sum_{t=0}^{q_k^{\max}} R^{(k)}(\nu-s_{a,k}t)\mu_{k\to a}(t) \tag{13}$$

for $k\in\{0,\dots,n_a-1\}$ and initial condition $R^{(0)}(0)=1$. The last step of the recursion gives us $m_{a\to i}(\nu)=R^{(n_a)}(\nu)$. The increase of performance is substantial since brute force integration involves a sum over $\prod_{l\in a\setminus i} q_l^{\max}$ terms, while iteration in Eq. refeq:iteration scales just as $\sum_{l\in a\setminus i} q_l^{\max}$. In some cases a faster convergence was met following a *refinement* strategy in the variable $q_i^{\max}$. Instead of choosing a large $q_i^{\max}$ from scratch and to solve the BP equations from random initial conditions, the simulation starts using small $q_i^{\max}$ that are increased after convergence. Each time $q_i^{\max}$ is increased the BP equations are solved starting from a function that fits the previous solution. With this strategy, networks as large as 40 metabolites and 120 reactions could be simulated in a couple minutes using a standard laptop. It should be noted that when the number n_a is large, the computations of $m_{a\to i}$ can be done substantially faster by means of discrete Fourier transforms, reducing the computation time of all messages $m_{a\to i}$ for $i\to a$ from the needed $n_a\times\sum_{l\in a\setminus i} q_l^{\max}$ operations to just around $2\times(\sum_{l\in a} q_l^{\max})\log(\sum_{l\in a} q_l^{\max})$.

Authors contributions

Authors equally contributed to this work.

Acknowledgements

AB was supported by Microsoft TCI grant. RM wants to thank the International Center for Theoretical Physics in Trieste and the Center for Molecular Immunology of La Habana, for their cordial hospitality during the completion of this work. We are also very grateful to Ginestra Bianconi, Michele Leone, Martin Weigt, and Riccardo Zecchina, for very interesting discussions, and in particular to Carlotta Martelli for sharing with us a human readable E-Coli data set.

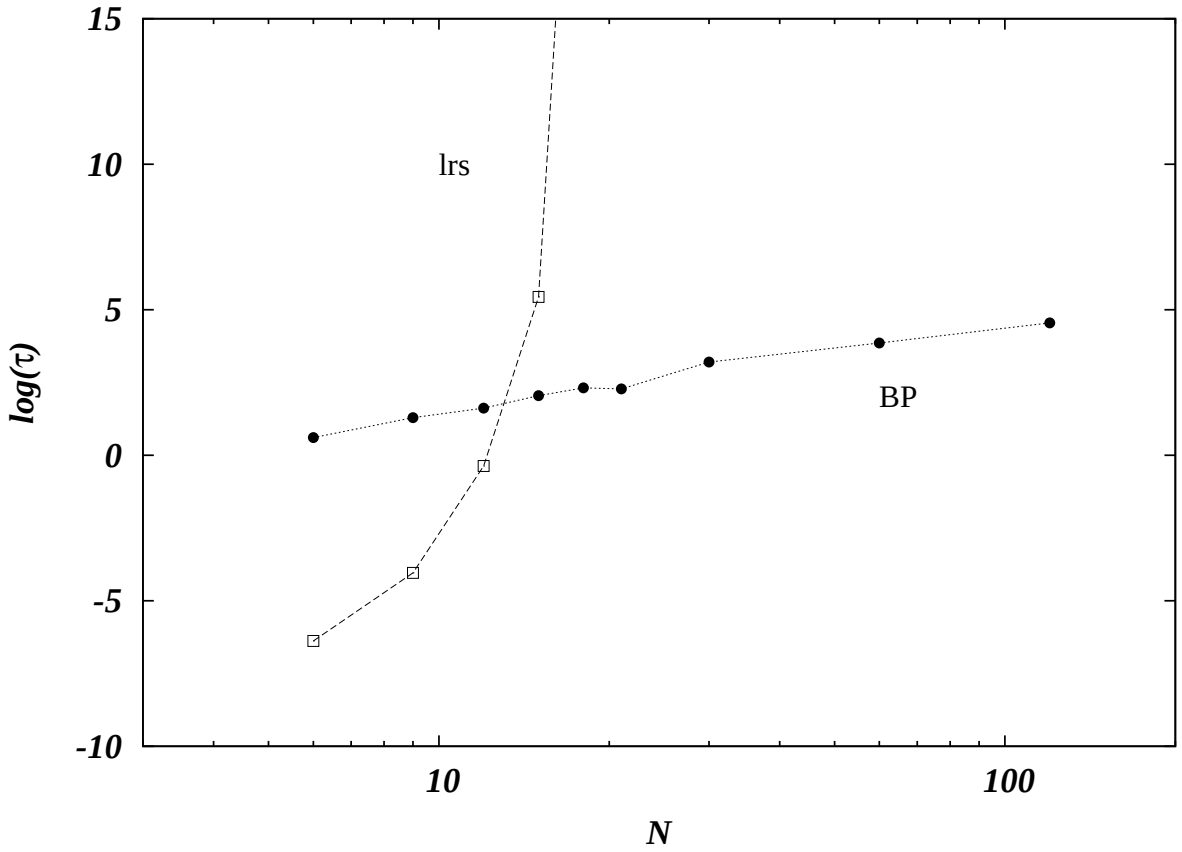
References

1. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**(6804):651–654.
2. Fell DA, Wagner A: **The small world of metabolism**. *Nature Biotechnology* 2000, **18**:1121–1122.
3. Z D, Qin ZS: **Structural comparison of metabolic networks in selected single cell organisms**. *BMC Bioinformatics* 2005, **6**(8).
4. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Res.* 2006, **34**:D354–7.
5. Ibarra AU, Edwards J, Palsson B: **Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth**. *Nature* 2002, **420**:186–189.
6. Varma A, Palsson B: **Metabolic Capabilities of Escherichia coli: I.Synthesis of biosynthetic precursors and cofactors**. *J. theor. Biol* 1993, **165**:477–502.
7. Segré D, Vitkup D, Church G: **Analysis of optimality in natural and perturbed metabolic networks**. *PNAS* 2002, **99**:15112–15117.

8. Wiback S, Famili I, Greenberg HJ, Palsson B: **Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space**. *J. Theor. Biol.* 2004, **228**:437–447.
9. **LRS package** [http://cgm.cs.mcgill.ca/~avis/C/lrs.html].
10. Avis D, Fukuda K: **A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra**. *Discrete Comput. Geom.* 1992, **8**(3):295–313.
11. Reed J, Vo T, Schilling C, Palsson B: **An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)**. *Genome Biology* 2003, **4**(9):R54, [http://genomebiology.com/2003/4/9/R54].
12. Almaas E, Kovacs B, Vicsek T, Oltval Z, Barabasi AL: **Global organization of metabolic fluxes in the bacterium Escherichia col**. *Nature* 2004, **427**:839–843.
13. A De Martino RM C Martelli, Castillo IP: **Von Neumann’s expanding model on random graphs**. *Journal of Statistical Mechanics: Theory and Experiment* 2007, **2007**(05):P05012, [http://stacks.iop.org/1742-5468/2007/P05012].
14. Bianconi G, Zecchina R: **Viable flux distribution in metabolic networks**. *Technical report* 2007, [http://aps.arxiv.org/pdf/0705.2816]. [ArXiv:cond-mat/0705.2816].
15. Beard D, Babson E, Curtis E, Qian H: **Thermodynamic constraints for biochemical networks**. *J. Theor. Biol.* 2004, **228**:327–333.
16. Dyer M, Frieze A: **On the complexity of computing the volume of a polyhedron**. *SIAM J. Comput.* 1988, **17**(5):967–97.
17. Khachiyan L: **Complexity of volume computation**. In *New trends in discrete and computational geometry*. Edited by Pach J, Springer-Verlag 1993:91–101.
18. Bëuler B, Enge A, Fukuda K: **Exact volume computation for convex polytopes: A practical study.** In *Polytopes–combinatorics and computation*. Edited by Ziegler GM, Kalai G, Birkhauser 2000:131–154.
19. Baxter B: *Exactly Solved Models in Statistical Mechanics*. London: Academic Press Inc. 1989.
20. Yedidia J, Freeman W, Weiss Y: **Generalized belief propagation**. In *Advances in Neural Information Processing Systems (NIPS) 13, Denver, CO*. Edited by press M 2001:772–778.
21. Kschischang FR, Frey BJ, Loeliger HA: **Factor graphs and the sum-product algorithm**. *Information Theory, IEEE Transactions on* 2001, **47**(2):498–519, [http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=910572].
22. Braunstein A, Mezard M, Zecchina R: **Survey propagation: An algorithm for satisfiability.** *Random Struct. Algorithms* 2005, **27**(2):201–226, [http://dblp.uni-trier.de/db/journals/rsa/rsa27.html#BraunsteinMZ05].
23. MacKay DJC: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press 2003, [http://www.cambridge.org/0521642981]. [Available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`].
24. Mezard M, Parisi G: **The Bethe lattice spin glass revisited**. *The European Physical Journal B* 2001, **20**:217, [http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0009418].
25. Mezard M, Parisi G: **The cavity method at zero temperature**. *J.Stat.Phys.* 2003, **111**:1, [http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0207121].

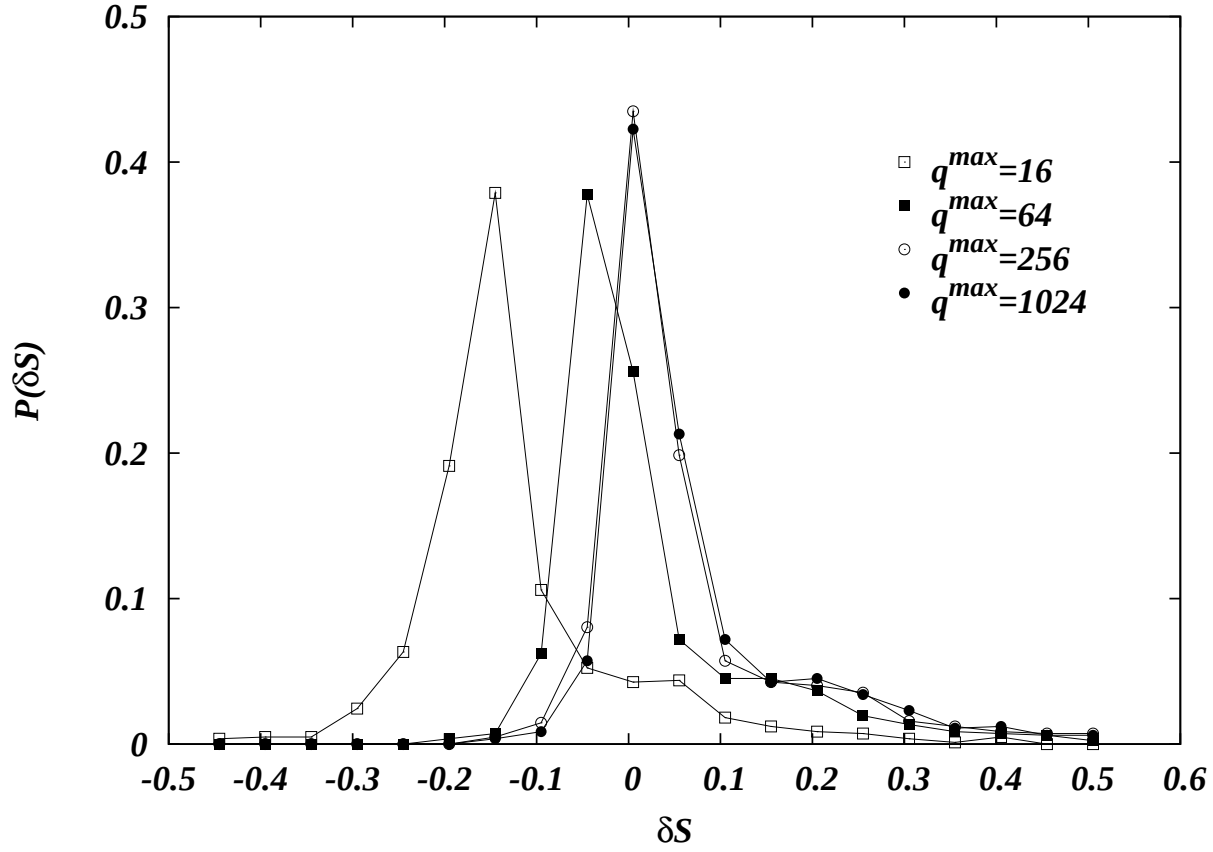
1 Figures

1.1 Figure 1.1 - Running time



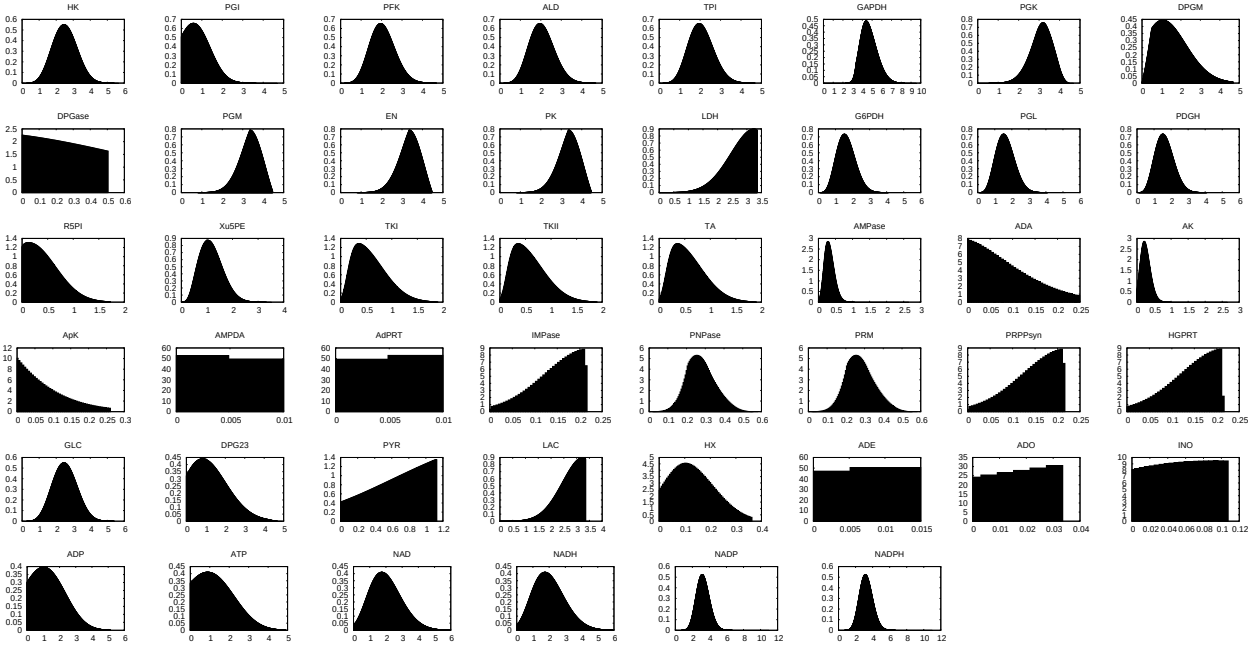
Logarithm of the running time vs. N for LRS algorithm, and BP algorithm. Averages were taken over 5000 realizations for the smaller lattices and 500 $N = 12$.

1.2 Figure 1.2 - Discrepance histogram



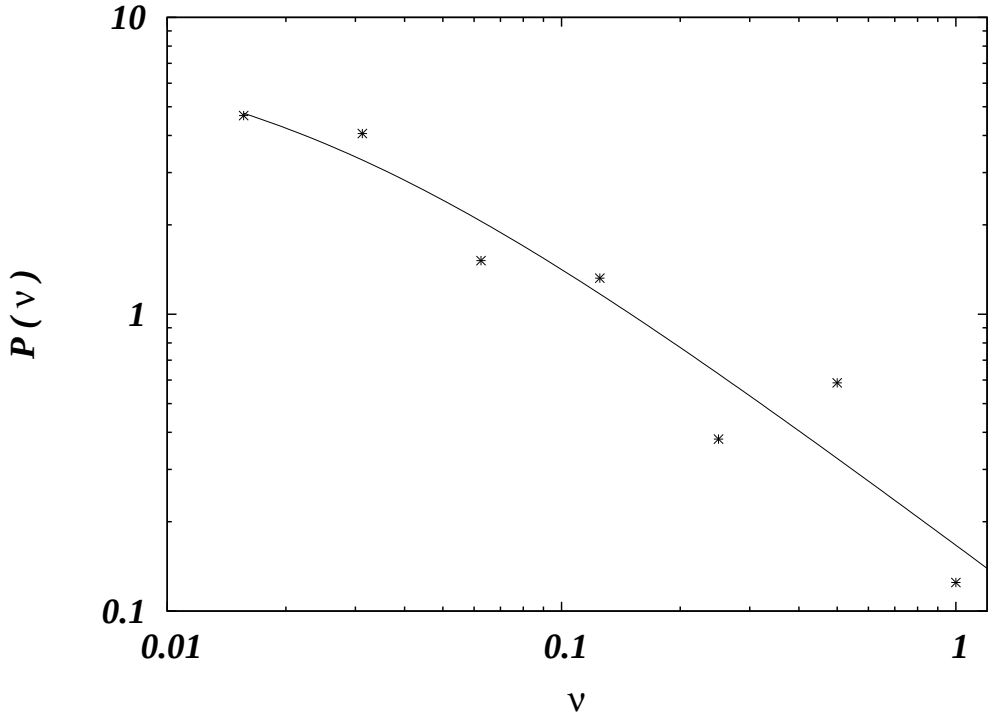
Histograms of $\delta_S = (S_{BP} - S_{LRS})/S_{LRS}$ over a set of 1000 realizations of the stoichiometric matrix. The three histograms are for $N = 12$, $M = 4$, and $K = 3$, at different value of $q^{\max} = 16, 64, 256, 1024$.

1.3 Figure 1.3 - Distribution of fluxes in Blood Cells



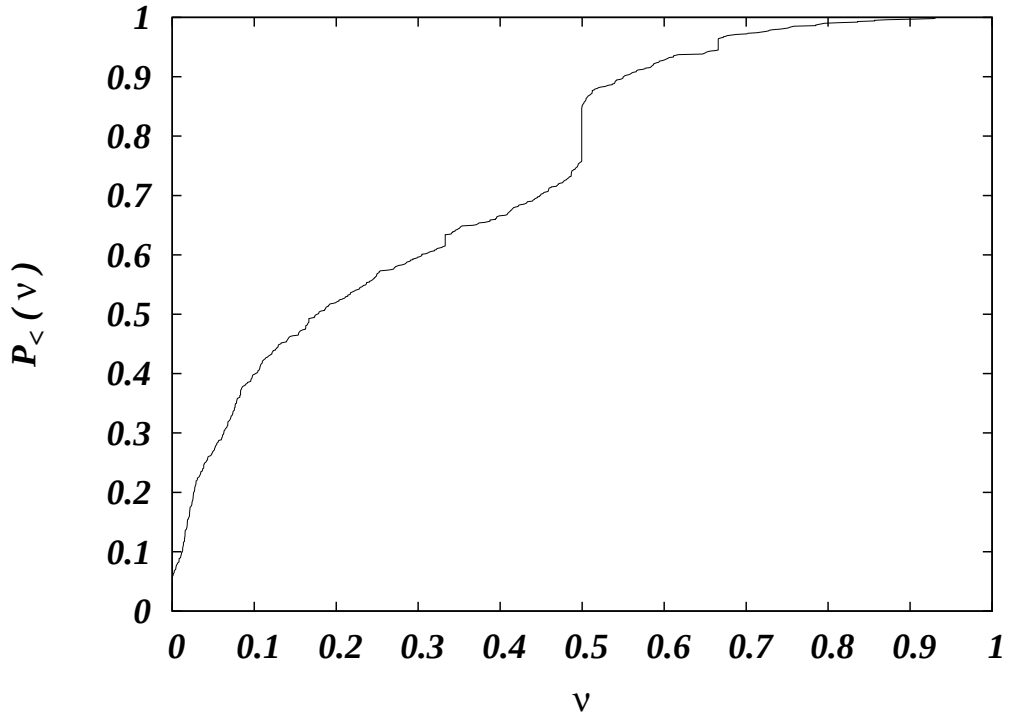
Distributions of the flux values for each reaction in the red blood cell network. They are arranged following the same sequence as in [8].

1.4 Figure 1.4 - Distribution of average fluxes



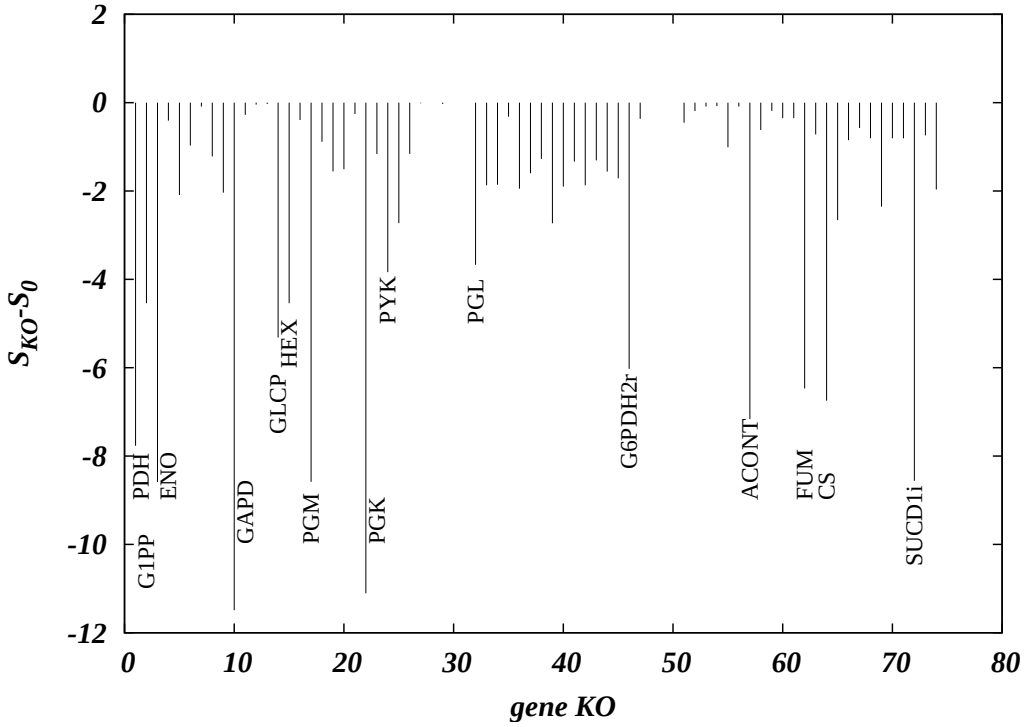
Distribution of average fluxes ($\bar{\nu}$) of the reactions for the E-Coli metabolism. $N = 1005$ and $M = 506$.

1.5 Figure 1.5 - Integrated distribution of average fluxes



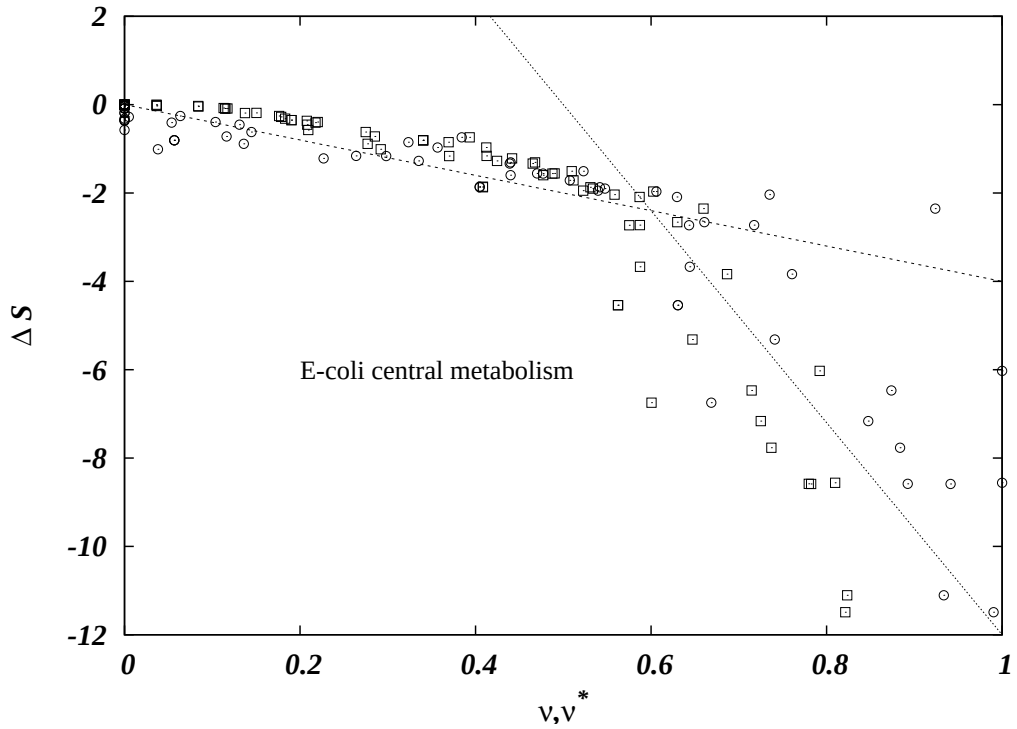
Integrated distribution of average fluxes ($\bar{\nu}$) of the reactions for the E-Coli metabolism. $N = 1005$ and $M = 506$. Note the jumps for $\bar{\nu} = 0.4$, $\bar{\nu} = 0.5$ and $\bar{\nu} = 0.6$

1.6 Figure 1.6 - Impact of knockout in E-Coli



Changes of S for different reaction knockouts in the central metabolic network of the E-Coli. Annotated reactions belong either to the glycolisis pathway or the Krebb cycle

1.7 Figure 1.7 - Correlation of impact and value



Correlations between the change in entropy ΔS after reaction knockouts and the average values ($\bar{\nu}$) and the most probable values (ν^*) of the fluxes in the central metabolism of the E-coli. The lines are guides to the eyes